From the Department of Medical Epidemiology and Biostatistics
Karolinska Institutet, Stockholm, Sweden

Novel statistical methods for genome-wide association summary statistics

Zheng Ning
宁铮

Karolinska Institutet

Stockholm 2020

All published papers reproduced with permission
Published by Karolinska Institutet
Printed by US-AB 2020

Typeset by the author using LaTeX 2$_\varepsilon$

ISBN 978-91-7831-887-2

Novel statistical methods for genome-wide association summary statistics

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Zheng Ning

Time and location: Friday 11 September 2020, kl 09.00 in the lecture hall Atrium, Nobels väg 12 B, Karolinska Institutet, Solna

Principal supervisor:
Assistant Professor Xia Shen
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Co-supervisor:
Professor Yudi Pawitan
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

Opponent:
Professor Jian Yang
University of Queensland
Institute for Molecular Bioscience

Examination board:
Professor Lars Rönnegård
Dalarna University
School of Technology and Business Studies

Professor Emeritus Hossein Jorjani
Swedish University of Agricultural Sciences
Department of Animal Breeding and Genetics

Associate Professor Åsa Johansson
Uppsala University
Department of Immunology, Genetics and Pathology

Professor Rebecka Jörnsten
University of Gothenburg / Chalmers
Division of Applied Mathematics and Statistics

Associate Professor Fredrik Wiklund
Karolinska Institutet
Department of Medical Epidemiology and Biostatistics

献给赐我生命的父母

To my parents, who gave me life

一杯敬明天，一杯敬过往

A toast to tomorrow, a toast to the past

Abstract

A general objective of genetic studies is to understand the genetic basis of complex traits such as height, body mass index (BMI), disease endpoints, etc. Such researches have been facilitated due to the completion of the human genome project and developments of high-throughput technologies. With the help of high-throughput genotyping and sequencing technologies, the information on millions of genetic markers can be measured for each individual.

The most widely used strategy to detect the associations between genetic variants and a complex trait is genome-wide association study (GWAS). Because the genetic architecture of most complex traits is highly polygenic, the signal to noise ratio is usually tiny. Thus, especially in human populations, GWAS often requires large samples to obtain sufficient power. Unfortunately, given the restrictions on sharing individual-level data, it is often not feasible to pool data from different cohorts. Despite that, in each cohort, it is possible to report and share GWAS summary statistics, such as sample sizes, allele frequencies, estimates of genetic effect sizes, and their standard errors for the genetic markers across the genome. Therefore one recent focus in statistical methodology development for genetic studies has been on meta-analysis techniques using summary-level data. The objective of this thesis is to develop novel statistical genetics methods based on GWAS summary statistics and to apply these methods to better understand the genetic architecture underlying complex traits.

In **Study I**, we developed a Selection Operator for JOint analyzing multiple SNPs (SOJO). We mathematically proved and empirically showed that the least absolute shrinkage and selection operator (LASSO) could be achieved using GWAS summary-level data. Compared to the stepwise selection procedures, SOJO performs better in variable selection. SOJO is useful for detecting additional variants with independent effects and assessing the magnitude of allelic heterogeneity within loci.

In **Study II**, we developed a High-Definition Likelihood (HDL) method to improve the accuracy in genetic correlation estimation using GWAS summary statistics. Compared to the state-of-the-art method LD Score regression (LDSC), HDL achieves higher statistical power to detect genetic correlations between phenotypes by fully accounting for linkage disequilibrium (LD) information across the genome.

In **Study III**, we introduced a four-level strategy for replication of loci detected by multi-trait GWAS methods. The four methods provide different degrees of replication strength, useful for providing additional evidence when a locus has been discovered and replicated by multivariate analysis of variance (MANOVA) or other multi-trait methods. The replication methods only require summary association statistics and are straightforward to be applied to multi-trait GWAS analyses.

In **Study IV**, using GWAS summary statistics, we developed a method named Genetic Correlation Contrast for Causality (G3C) as a more robust test for the existence and direction of causal relationships between phenotypes. In contrast to Mendelian Randomization (MR), G3C does not rely on the assumption of no horizontal pleiotropy. G3C takes full advantage of genome-wide genetic association data and account for underlying genetic correlations between complex traits.

List of scientific studies

I. Zheng Ning, Youngjo Lee, Peter K. Joshi, James F. Wilson, Yudi Pawitan, and Xia Shen
A selection operator for summary association statistics reveals allelic heterogeneity of complex traits
The American Journal of Human Genetics **101**: 903–912.

II. Zheng Ning, Yudi Pawitan, and Xia Shen
High-definition likelihood inference of genetic correlations across human complex traits
Nature Genetics **52**: 859–864.

III. Zheng Ning, Yakov A. Tsepilov, Sodbo Zh. Sharapov, Zhipeng Wang, Alexander K. Grishenko, Xiao Feng, Masoud Shirali, Peter K. Joshi, James F. Wilson, Yudi Pawitan, Chris S. Haley, Yurii S. Aulchenko, and Xia Shen
Nontrivial replication of loci detected by multi-trait methods
Submitted

IV. Zheng Ning, Peter K. Joshi, Youngjo Lee, James F. Wilson, Yudi Pawitan, and Xia Shen
Inferring causation from heterogeneity in genetic correlations of complex traits
Manuscript

The articles will be referred to in the text by their Roman numerals, and are reproduced in full at the end of the thesis.

Contents

List of abbreviations

AH	Allelic heterogeneity
BMI	Body mass index
CAD	Coronary artery disease
CP	Cross-phenotype
eQTL	Expression quantitative trait loci
G3C	Genetic correlation contrast for causality
GCTA	Genome-wide complex trait analysis
GCTA-COJO	Conditional and joint multiple-SNP analysis
GIANT	The Genetic Investigation of ANthropometric Traits
GREML	Genome-based eestricted maximum likelihood
GRM	Genomic relationship matrix
GWAS	Genome-wide association study
HDL	High-definition likelihood
IV	Instrumental variable
LASSO	The least absolute shrinkage and selection operator
LD	Linkage disequilibrium
LDSC	LD score regression
LMM	Linear mixed model
MANOVA	Multivariate analysis of variance
MR	Mendelian randomization
RCT	Randomized controlled trial
REML	Restricted maximum likelihood
SNP	Single-nucleotide polymorphism
SOJO	Selection operator for jointly analyzing multiple variants
UKB	UK Biobank

Chapter 1

GWAS and GWAS meta-analysis

GWAS

In terms of mapping genes and genetic variants affecting Mendelian traits, linkage analysis has been successful (1). However, because of the relatively low power and resolution for gene mapping, linkage analysis has had limited achievements when it comes to complex traits or common diseases (2). In 1996, Risch and Merikangas (3) discussed the potential power boosting by using association analysis instead of linkage analysis. This imaginative design soon became reality because of the technological advances and biobanks established around the world. The first discovery based on GWAS was reported in 2005 (4), where a single-nucleotide polymorphism (SNP) was found to be associated with age-related macular degeneration. The study only included 96 cases and 50 controls, and 116,204 SNPs. As of 2017, about 10,000 strong associations ($P < 5 \times 10^{-8}$) between genetics variants and complex traits have been reported (5). According to the NHGRI-EBI GWAS Catalog (6), as of 14 July 2020, 4,628 publications and 189,811 associations have been recorded. The abundant results generated in GWASs facilitate geneticists' understanding of the genetic architecture underlying common diseases and complex traits.

According to GWAS results (7; 8; 9), most studied complex traits are highly polygenic, which means many polymorphisms in many genes contribute to the genetic components of the complex traits together. As a consequence, each single genetic variant can usually only explain a small fraction of phenotypic variance. On the other hand, this also suggests that larger sample size will lead to more discoveries, which has been happening during the past years. Take human height as an example, in 2008, 54 SNPs had been identified and jointly explained only ~5% heritability. By then, the sample size was ~63,000 (10). When the sample size was enlarged to ~250,000 in 2014, the number of identified associated SNPs increased to ~700, explaining ~20% of the trait's heritability (7). As more and more new GWAS data are still being collected, we can foresee more identified SNPs for each complex trait in the next few years.

The numerous GWAS results suggest another implication: Widespread pleiotropy for complex diseases and traits. For example, some GWAS indicated that the same SNP or

gene set could be significantly associated with different traits, given that the phenotypes are measured in independent samples (11). More evidence in auto-immune disease studies implies that at some loci, the same causal variants can drive the observed association across different diseases (12; 13).

GWAS has led to many scientific discoveries. However, many people have voiced their concerns about GWAS. A major concern is the biological implication of the GWAS discoveries (14). By its design, an association reported by GWAS does not directly yield a causal variant or a specific gene target of the underlying molecular mechanism (See chapter 3). Nevertheless, GWAS discoveries play an important role as the first step of the discovery pipeline. Based on GWAS signals, scientists have found several target genes by performing subsequent functional experiments. For example, the *FTO* region harbours the strongest genetic association with BMI (8). Following this signal, researchers discovered the underlying pathway between *FTO* region and *IRX3*; and *IRX3* was suggested as a body mass and composition regulator (15).

GWAS meta-analysis

GWAS meta-analysis can be used to combine results from multiple GWAS in independent cohorts. By increasing sample size, GWAS meta-analysis boosts statistical power to discover new genetic loci for common diseases and traits (16). Given the limited power in single-cohort GWAS and the difficulty of sharing individual-level data, GWAS meta-analysis based on summary-level data has become a popular approach, especially in human genetics. In the past few years, most discovered associations were from large-scale GWAS meta-analyses (17).

The most popular meta-analysis approach for synthesizing GWAS data is the fixed effects model with inverse variance weights (17). Assuming the true effect of each causal variant is the same across studies, the pooled estimator of effect size is

$$\hat{\beta}_M = \frac{\sum_j w_j \hat{\beta}_j}{\sum_j w_j},$$

with estimated variance of

$$\mathrm{Var}(\hat{\beta}_M) = \frac{1}{\sum_j w_j},$$

where the weight

$$w_j = \frac{1}{\mathrm{Var}(\hat{\beta}_j)}.$$

In some cases, the between-study heterogeneity is non-negligible (18). To take this heterogeneity into account, random effects model based on DerSimonian and Laird estimator

can be used (19). The weight is calculated as

$$w_j^R = \frac{1}{\frac{1}{w_j} + \hat{\tau}^2}, \text{where } \hat{\tau}^2 = \frac{Q - (k-1)}{\sum_j w_j - (\frac{\sum_j w_j^2}{\sum_j w_j})}. \tag{1.1}$$

The Q in (1.1) is Cochran's Q statistic for measuring the amount of between-study heterogeneity, which is given by

$$Q = \sum_j w_j (\hat{\beta}_j - \hat{\beta}_M)^2.$$

Comparing to the random effects model, the fixed effects model had better discovery power. Therefore the fixed effects model is more appropriate when the aim is to make discoveries (20). On the other hand, the random effects model is preferable when the generalizability of results is of interest, e.g., in prediction problems (18).

Chapter 2

Aims of the thesis

Summary statistics from GWAS or GWAS meta-analysis are valuable resources. Particularly, they are sufficient statistics for many statistical genetics methods. With sufficient information, methods based on summary-level data can perform similar to those based on individual-level data. In many cases, summary-level methods are preferable because (i) they allow a much larger sample size, which boosts statistical power; and (ii) they are more computationally efficient. The overall aim of this thesis is to better analyze the genetic architecture underlying complex traits by developing and implementing statistical methods based on GWAS summary statistics. The specific aims of individual studies are:

I. To develop and implement the least absolute shrinkage and selection operator (LASSO) using GWAS summary-level data for better fine-mapping.

II. To improve genetic correlation estimation using GWAS summary statistics based on a full likelihood method.

III. To introduce a more rigorous and meaningful replication strategy for loci detected by multi-trait GWAS methods, based on GWAS summary statistics.

IV. To provide a test for the existence and direction of causal relationships between phenotypes, accounting for genetic correlations based on GWAS summary statistics.

Chapter 3

Fine-mapping and allelic heterogeneity

Fine-mapping

The number of SNPs considered in GWAS is small relative to the total number of existing SNPs in the genome. As a result, the causal variants, i.e., the genetic variants causing the trait's variation, are usually not included in GWAS. Despite that, GWAS has successfully identified SNPs associated with complex traits across the whole genome. Why does GWAS work even though the causal variants may not be included? The answer is the correlation or linkage disequilibrium (LD) between the SNPs in GWAS and the actual causal variants. If a SNP is correlated with a causal variant, the SNP will show some signal in GWAS. In this case, the SNP is called a tag SNP or tagging SNP because it tags the true causal variant (21). In other words, most SNPs suggested by GWAS are tag SNPs instead of causal variants themselves (22). At this point, we need fine-mapping methods to refine the genomic localization of causal variants.

A typical flow of the fine-mapping process is shown in Fig 3.1. Because the regional LD structure can be complicated, fine-mapping becomes challenging, especially when a locus harbors multiple causal variants. Early efforts were put on so-called LD clumping, which filters out the SNPs correlated to the lead SNP. As a simple fix, LD clumping does not account for the joint effects of the SNPs. Therefore SNPs can be filtered out barely due to their mild LD with the lead SNP, even though they may tag independent causal variants. To solve this issue, GCTA-COJO (23) and Bayesian methods (24) were developed to model multiple tag SNPs jointly. A short introduction to these methods will be provided in the later sections in this chapter.

After identifying potential SNPs tagging independent causal variants, how can we be confident that they are real discoveries? There is a common hypothesis that a causal variant affects traits by regulating the expression of a nearby gene in some tissues. Therefore, besides replicating the SNPs in another cohort, we can investigate whether they are expression quantitative trait loci (eQTL) (25). If a SNP is significant in both GWAS and eQTL analysis, then the colocalization information makes the pathway and the discovered SNP more convincing (26). However, in many cases, GWAS loci are not necessarily strong

eQTL (25). Instead, they may regulate genes in other ways, such as alternative splicing (27).

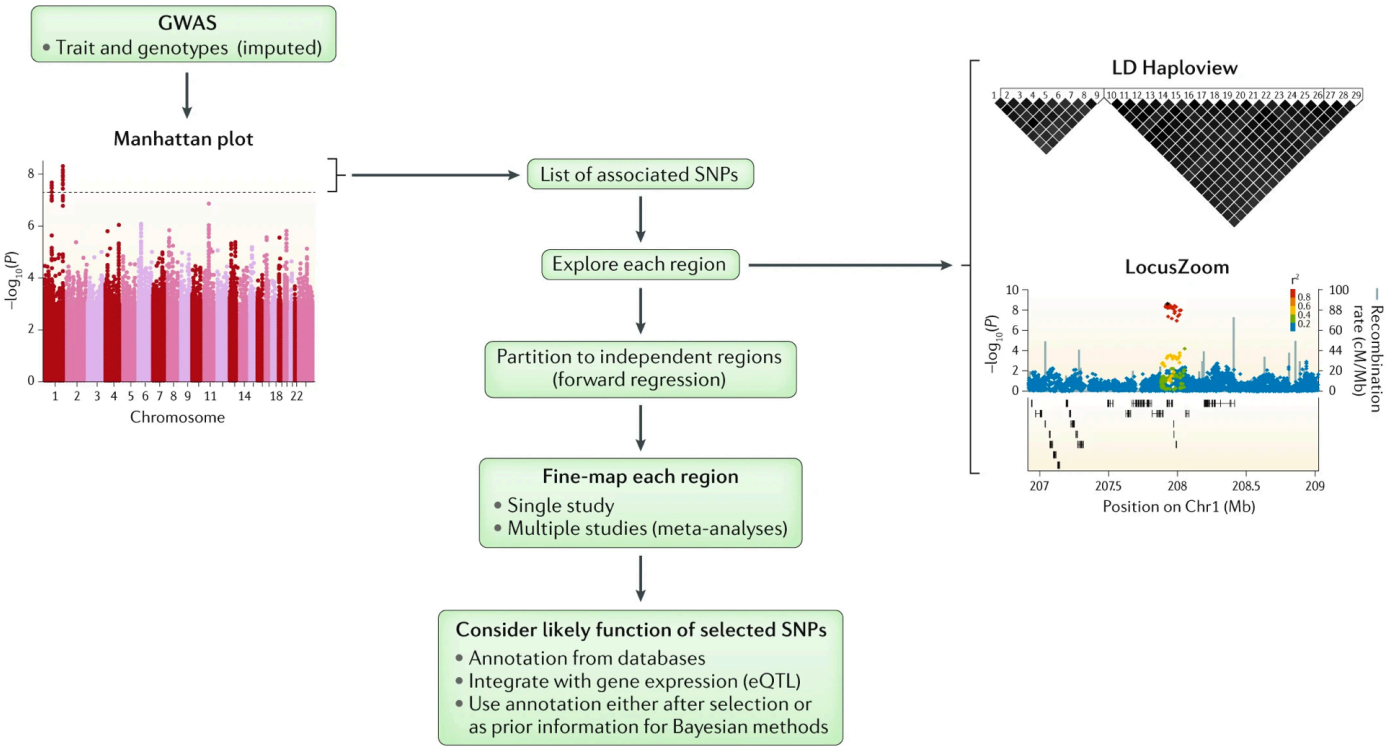

Figure 3.1: A typical flow of fine-mapping process from GWAS to annotation of SNPs selected from fine-mapping. The figure is reprinted from Schaid et al. (28) with permission from Springer Nature.

Allelic heterogeneity

Allelic heterogeneity (AH) is the phenomenon when there are multiple causal variants at the same locus for a trait. Different from polygenicity, which measures the spread of genetic effects across the genome, AH evaluates the spread of genetic effects within each locus. Hence AH together with polygenicity measure the complexity of a trait. AH of some Mendelian traits such as cystic fibrosis has been well studied (29). As a contrast, the extent of AH for complex traits is unclear. Although AH is reported to be present in various complex diseases (30), more evidence and quantitative results are needed.

GCTA-COJO

GCTA-COJO represents conditional and joint multiple-SNP analysis. Based on GWAS meta-analysis summary statistics, GCTA-COJO performs a secondary association scan conditioning on the discovered top variants. For an individual, assuming a quantitative trait y is potentially affected by a group of genetic variants $X_1, ..., X_p$ and a multi-variant

linear model

$$y = X\boldsymbol{\beta} + e.$$

The joint effects of multiple SNPs estimated by the least-squares are

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}, \text{ and } \mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma_J^2(\mathbf{X}^T\mathbf{X})^{-1},$$

where $\mathbf{X}^T\mathbf{X}$ can be estimated from the LD structure of a reference sample, such as a subcohort of the GWAS meta-analysis; and $\mathbf{X}^T\boldsymbol{y}$ can be estimated from GWAS meta-analysis summary-level data. Then the joint p-values can be approximated. Conditional p-values are derived similarly. Based on conditional p-values, a genome-wide stepwise selection is performed. When there is no more variant can be added or removed in the model, the joint effects and p-values are reported. GCTA-COJO has been widely applied in global consortia such as GIANT and DIAGRAM (7; 8; 9; 31), where additional associations in LD at the same loci were successfully detected.

The stepwise selection, which is implemented in GCTA-COJO, has been used as a standard variable selection procedure in many fields. After measuring variables, researchers frequently perform multiple linear regression analyses to select relevant variable subsets and derive models. The most straightforward technique is to check all possible variable subsets and select the best one. However, considering the number of different combinations of variables, the amount of computation can be intractable even when efficient algorithms are implemented (32). Therefore stepwise methods such as forward selection and backward elimination are used. But in the forward selection, a formerly selected variable may become unimportant as other variables enter the model; similarly, in backward elimination, a former deleted variable may become important as other variables are removed from the model. To deal with these problems, the stepwise selection procedure was developed (33). The stepwise selection is a combination of the forward and backward selection: Variables are added to the model one at a time; and at each step, backward elimination is performed to see whether variables can be removed from the model.

However, the literature suggests that the stepwise selection is very likely to remove some useful variables which are in mild correlation with selected predictors (34). This indicates that GCTA-COJO may miss some informative tag SNPs due to their mild LD with selected variants. By setting a less stringent significance threshold, more variants can be selected. But there will be a substantial risk of overfitting for the model generated by the stepwise selection, especially when the model includes many variants (35).

LASSO

Comparing to the stepwise selection, literature suggests that simultaneous modeling of multiple predictors with penalization performs better in variable selection (36). In 1996, Tibshirani introduced the least absolute shrinkage and selection operator (LASSO) (37), which has been applied across various disciplines since then (38; 39). Besides the square

loss function $\frac{1}{2}\|\mathbf{y}-\mathbf{X}\hat{\beta}\|_2^2$, LASSO takes the ℓ_1-norm regularization $\|\hat{\beta}\|_1$ into account and solves

$$\min_{\hat{\beta}\in R^p} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1,$$

where the tuning parameter $\lambda \geq 0$. The ℓ_1 term serves as a penalization: the larger λ is, the larger the coefficients are penalized. Because of the penalization term, large coefficients are allowed only when they lead to substantially better fit. Consequently, LASSO can perform variable selection, and achieves better interpretability and prediction accuracy (37). Besides, by performing a more reasonable bias-variance trade-off with shrinkage, overfitting problem is alleviated in LASSO. Therefore LASSO allows more informative predictors in the model without serious overfitting. Owing to the LARS algorithm (34) and regularization path algorithm (40), solving LASSO becomes computationally fast. In recent decades, LASSO has been used in genetics researches to select variants (41) and build prediction model (42).

In **Study I**, we developed a selection operator for jointly analyzing multiple variants (SOJO). SOJO performed LASSO regression within each mapped locus by using summary association statistics (43). SOJO has been shown to be more powerful than GCTA-COJO in terms of both discovery and prediction.

Bayesian methods

Many Bayesian methods have been proposed for fine-mapping (44; 24; 45). The typical steps of the Bayesian methods are (i) assuming a prior distribution of the true effect sizes; (ii) letting each SNP has two possible states: included or not included; (iii) computing the posterior probabilities for the different models; (iv) computing the posterior inclusion probability of each SNP and (v) selecting SNPs based on their posterior inclusion probabilities. When the true causal variants are included in GWAS, comparing to conditional analysis such as GCTA-COJO, Bayesian methods are shown to have a higher probability of selecting the actual causal variants (24). However, it is worthy to note that in most cases, the actual causal variants are not included in GWAS. Another disadvantage of Bayesian methods is its heavy computational burden (24; 45).

Chapter 4

Genetic correlation

One genetic variant or locus can be causal for more than one trait. This phenomenon is termed as pleiotropy. A real pleiotropy between two traits can be due to various biological reasons (46), such as causation or sharing a common biological process. One step further, genetic correlation evaluates the alignment of pleiotropy across the whole genome. An extensive existence of genetic correlations between complex traits and diseases has been reported (47; 13). Genetic correlation can help understand the shared biology underlying complex traits. Following significant genetic correlations between traits, different methods can be used to (i) perform causal inference (48), (ii) improve power to detect new associations by modelling multiple traits simultaneously (See chapter 5), (iii) improve genetic risk prediction (49) and (iv) better understand the relationships between traits by further modelling genetic correlated traits (50).

The mathematical definition of genetic correlation is based on the general quantitative genetic model. For each individual, the phenotype of trait 1 (y_1) is the sum of a genetic value (g_1) and a residual (e_1):

$$y_1 = g_1 + e_1.$$

Similarly for the phenotype of trait 2 (y_2):

$$y_2 = g_2 + e_2.$$

Denoting the genetic variance of the two traits as $\sigma^2_{g_1}$ and $\sigma^2_{g_2}$ respectively, and the covariance of the genetic values as σ_{g_1,g_2}, the genetic correlation (r_g) is defined as

$$r_g = \frac{\sigma_{g_1,g_2}}{\sqrt{\sigma^2_{g_1}\sigma^2_{g_2}}}.$$

If the phenotypic variances of the two traits are standardised to one, and the genetic values are limited to only additive effects, then the genetic variances $\sigma^2_{g_1}$ and $\sigma^2_{g_2}$ become narrow-sense heritabilities h^2_1 and h^2_2, and the genetic covariance σ_{g_1,g_2} is represented by coheritability h_{12}.

Conventionally, to estimate h^2_1, h^2_2 and h_{12}, a bivariate linear mixed model (LMM) is

applied to a cohort with pedigree information. Then the parameters can be estimated using restricted maximum likelihood (REML) (51). However, given the difficulty of collecting complete family data, the power of family-based methods is often limited by the small sample size. Large sample size can be obtained from national registries (52), which is available only in a few countries. Besides, the shared environmental factors within families may bias the estimates.

Bivariate GREML

Thanks to the advance of genomic technology, more and more large cohorts with genotypes and phenotypes of independent individuals have become available. To estimate h_1^2, h_2^2 and h_{12} based on data from independent individuals, Yang et al. developed genome-based REML (GREML) in their genome-wide complex trait analysis (GCTA) software (53). Similar to REML for pedigree data, GREML also uses an LMM. In REML for pedigree data, the variance-covariance structure is derived from a kinship matrix; while in GREML, the kinship matrix is replaced by a genomic relationship matrix (GRM) derived from genome-wide SNP genotype data. Conceptually, all "independent individuals" are distant relatives. The GRM is a good approximation of the kinship between distant relatives. Bivariate GREML (54) can estimate h_1^2, h_2^2, h_{12} and r_g simultaneously.

GREML has been widely used since its inception (55; 56; 57). However, REML itself is computationally intensive. Although various algorithms have been suggested to speed it up (58; 59; 60), REML is still computationally challenging when the number of individuals is very large such as in UK Biobank (UKB).

LD Score regression

In GWAS, the χ^2 association statistic for a SNP is driven by the effects of all SNPs tagged by this SNP (61). Therefore, for polygenic traits, SNPs having higher LD with other SNPs tend to have higher χ^2 statistics on average. Motivated by this fact, LD Score regression (LDSC) was developed (62). Because LDSC only relies on GWAS summary statistics, it is computationally very efficient and convenient to use.

In LDSC, under the same LMM as in GCTA, the expected χ^2 statistic for variant j is:

$$\mathrm{E}\left[\chi_j^2 \mid l_j\right] = \frac{Nh^2}{M} l_j + Na + 1, \tag{4.1}$$

where $l_j = \sum_k r_{jk}^2$ is called as the *LD Score* of variant j, which measures the amount of genetic variation captured by variant j; N is the sample size; M is the number of variants in the polygenic model and a is the contribution of confounding biases such as population stratification. When there are two traits, and the confounding biases are absent, equation

(4.1) can be extended to

$$\mathrm{E}\left[z_{1j}z_{2j} \mid l_j\right] = \frac{\sqrt{N_1N_2}h_{12}}{M}l_j + \frac{\rho N_s}{\sqrt{N_1N_2}}, \tag{4.2}$$

where h_{12} is the genetic covariance; N_i is the sample size for study i; N_s is the number of overlapping individuals; and ρ is the phenotypic correlation among the N_s overlapping samples. Then genetic correlation can be computed as $r_g = h_{12}/\sqrt{h_1^2h_2^2}$, where h_i^2 is the heritability of trait i estimated from study i. This bivariate LDSC (47) has been widely used in genetics and epidemiology to estimate genetic correlations between various traits and diseases (63; 64). A centralized database and web interface has been developed to integrate and get LDSC results for hundreds of traits/diseases (65).

As a ratio of genetic covariance and heritabilities, r_g is a robust estimate because the biases on the numerator and the denominator often cancel out (66). r_g from LDSC has been shown to be robust against multiple factors, such as model misspecification (67), scale transformation (47), ascertainment of cases and strong environmental factors (68). However, it is worthy to note that the efficiency of LDSC is lower than that of GREML (47), even when the GWAS sample and the reference sample matched.

In **Study II**, we introduced that LDSC only partially uses LD information, which is an essential source generating the efficiency gap between LDSC and GREML. We proved that the LD matrix determined not only the variance of the single SNP test statistic but also the whole variance-covariance matrix of the test statistics. Therefore (4.1) and (4.2) can be extended to

$$\begin{aligned}\mathrm{Cov}\left[\mathbf{z}_i\right] &= \frac{N_ih_i^2}{M}\mathbf{L} + \mathbf{R}\\ \mathrm{Cov}\left[\mathbf{z}_1, \mathbf{z}_2\right] &= \frac{\sqrt{N_1N_2}h_{12}}{M}\mathbf{L} + \frac{N_s(h_{12}+\rho)}{\sqrt{N_1N_2}}\mathbf{R}\end{aligned}$$

where $\mathbf{z}_i$ is the Z score vector of the M SNPs from study i of trait i, $\mathbf{R}$ is the LD matrix of the M SNPs, $\mathbf{L} := \mathbf{R}'\mathbf{R}$ is the *LD score matrix*.

To fully use LD information, we developed high-definition likelihood (HDL) to estimate genetic correlations as a full likelihood-based method using summary-level data (69). Comparing to LDSC, HDL estimates genetic correlations more accurately.

Chapter 5

Multi-trait methods

The small genetic effects of SNPs on complex traits limit the discovery power in most GWAS. One approach to improve statistical power is jointly modeling multiple correlated phenotypes (70). At the early stage, most multi-trait tools were based on individual-level data. For example, the `--multivariate` module of PLINK implements canonical correlation analysis to identify the association between each SNP and linear combinations of phenotypes (71); Combined-PC (72) performed a principal components analysis on the phenotype data to improve statistical power.

In recent years, multi-trait methods based on GWAS summary-level data became popular. Many such methods have been developed and shown their benefits in terms of boosting discovery power (73; 74; 75; 76). For example, Stephens (77) outlined a unified multivariate analysis framework based on Bayesian model comparisons; Zhu et al. (78) introduced two test statistics S_{Hom} and S_{Het} to improve statistical power under different assumptions of effect sizes, using which seven additional loci were suggested by jointly analyzing the summary statistics of three traits from the GIANT consortium (79); MTAG (Multi-trait analysis of GWAS) was developed to integrate GWAS summary results of multiple related traits and improve the inference in each single-trait GWAS (49).

A simulation study (80) demonstrated that the statistical powers of most methods are similar to the power of the standard multivariate analysis of variance (MANOVA). It has been shown that MANOVA can be performed using summary statistics (77; 78). If we denote the true marginal effects of a SNP on k traits by $\boldsymbol{\beta}$, then the null hypothesis in MANOVA is $H_0 : \boldsymbol{\beta} = \mathbf{0}$. Let $\mathbf{t} = [t_1, ..., t_k]'$ be the vector of single-trait t-test statistics from association tests between the genotypes $\mathbf{g}$ of a single SNP and the k phenotypes, and $\mathbf{R}^* \equiv \text{cor}(\mathbf{t} \mid \boldsymbol{\beta}) = \text{Var}(\mathbf{t} \mid \boldsymbol{\beta})$. If $\mathbf{R}^*$ is available, the test statistic

$$T^2 = \mathbf{t}'\mathbf{R}^{*-1}\mathbf{t},$$

which asymptotically follows a χ^2 distribution with k degrees of freedom under the null hypothesis. In practice, an unbiased estimate of $\mathbf{R}^*$ can be obtained by selecting a large number of independent variants from the GWAS summary statistics and calculating their correlation coefficients (77; 78).

Although most multi-trait methods can boost discovery power, the replication strategy has yet to be agreed upon. When a SNP is discovered in multi-trait analysis, a commonly used approach for replication is to replicate the associations trait-by-trait in a replication sample (81; 82). However, there are two disadvantages of such "univariate" replication: (i) when the number of tested traits is large, multiple testing arises when determining the replication significance threshold; (ii) univariate replication does not normally account for phenotypic correlations between the tested traits, which generates conservative significance threshold after correction for multiple testing. Another straightforward way for replication is to directly perform the multi-trait test in a replication sample and see whether the overall association (omnibus p-value) is significant (83; 81; 84). Although this strategy provides a unified test statistic, the consistency of the multiple genetic effects between the discovery and replication samples is usually overlooked. Even if the genetic effect sizes and directions are distinct between the discovery and replication samples, the multivariate replication test may still declare significance.

In **Study III**, we introduced a four-level replication strategy for replication of loci detected by multi-trait methods. The four methods provide different degrees of replication strength, useful for extra evidence when a locus has been replicated by MANOVA or other multi-trait methods. The replication methods only require summary association statistics, straightforward to be applied to multi-trait GWAS analyses.

Pleiotropy and cross-phenotype association

Pleiotropy is normally defined as the same locus affecting multiple phenotypes (73); while cross-phenotype (CP) association occurs when a locus or genetic variant is statistically associated with multiple phenotypes, regardless of the causes of the observed associations (Fig 5.1). Therefore CP association is necessary but not sufficient for pleiotropy with any causal implication.

Nevertheless, practically, it is more promising to discover pleiotropy from loci with reliable and validated CP associations. Therefore, in **Study III**, we proposed the correlation methods to replicate CP associations in a different cohort. This out-of-sample replication can help rule out spurious pleiotropy. Spurious pleiotropy often arises in high linkage disequilibrium (LD) region, where a SNP is more likely to tag multiple causal variants located in different genes with distinct functions, such as in the major histocompatibility complex region (85) and the immunoglobulin heavy chain (*IGH*) locus (12). Given the LD difference between discovery and replication samples, spurious pleiotropy is expected to be rare among properly validated CP associations.

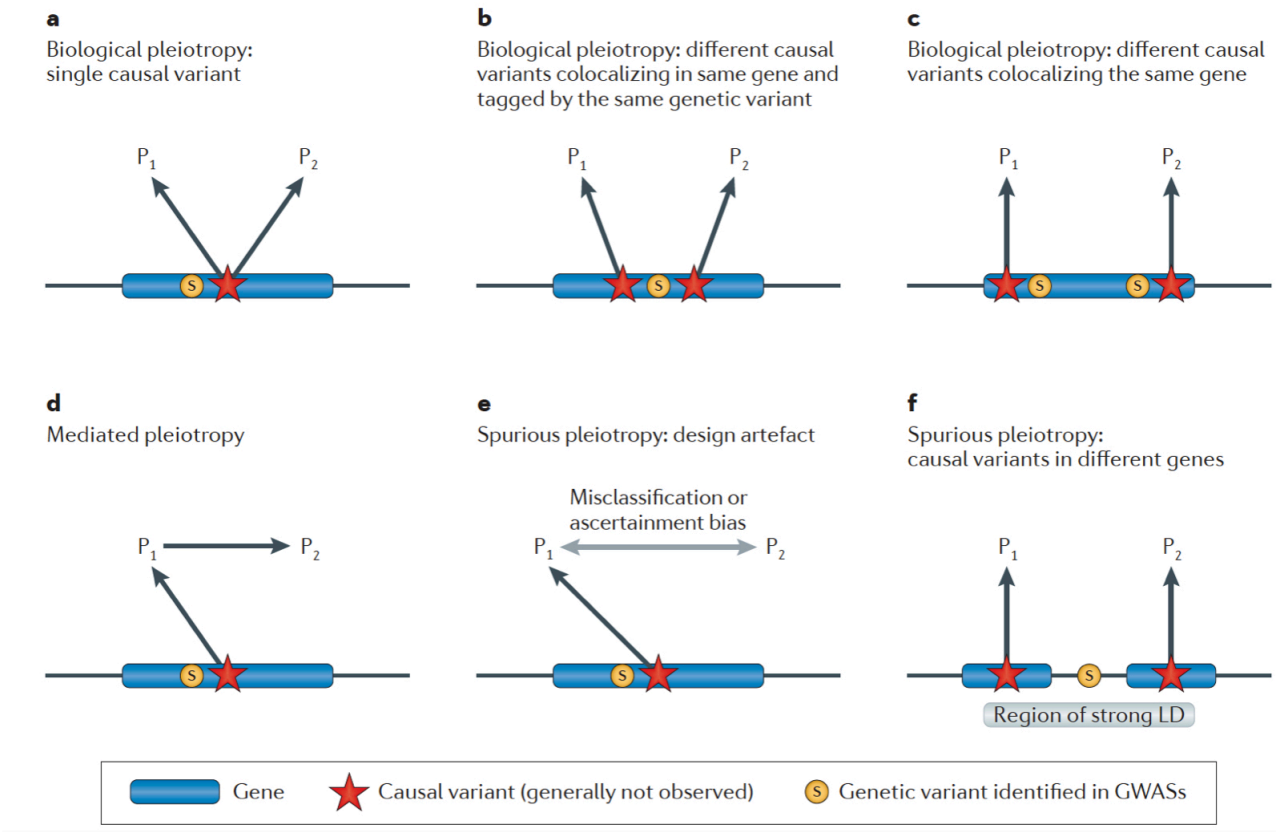

Figure 5.1: Different types of pleiotropy. The figure is reprinted from Solovieff et al. (73) with permission from Springer Nature.

Chapter 6

Causal inference using genetic data

A key objective of epidemiology and biomedical science is to understand the underlying causalities between phenotypes. A real causality can serve as the foundation for public policies or the target of drugs. However, it is challenging to infer causality from observational data. A primary issue is confounding, which introduces an association between exposure and outcome in the absence of a causal relationship. For example, when we study whether vitamin E is a protective factor for coronary artery disease (CAD), diet is a confounder because it affects both vitamin E intake and CAD risk (86).

To tackle the confounding problem properly, a precise definition of causality is needed. The most commonly used way to define causality is the counterfactual approach (87): An individual is exposed in one world, while unexposed in another identical world. The only difference between the two worlds is the exposure status of the individual. In this hypothetical setting, a causal effect can be defined as the outcome difference of the individual between the two worlds. Although such a scenario is impossible in reality, a consistent estimate of causal effect can be attained if the exposure status in the real world is independent to the potential counterfactual outcomes, which is termed as exchangeability. In other words, exchangeability means that there is no systematical difference between the exposed and non-exposed groups. Randomized controlled trial (RCT) ensures exchangeability by randomly split individuals into the exposed or non-exposed groups. Therefore RCT is often regarded as the gold standard for causal inference.

Although RCT is powerful, it is usually expensive and sometimes infeasible or unethical to conduct. Thus causal inference based on observational data is often necessary and in great need. Theoretically, if all confounders are controlled for, then conditional exchangeability can be reached, where exchangeability holds in each stratum of the combination of confounders. In this case, the causal effect can be consistently estimated. However, in reality, we can never rule out the existence of unobserved confounders and incorrectly controlled confounders. Different confounder controlling is an important source that generates inconsistent results across observational studies.

Mendelian Randomization

The emerging genetic data and genetically informed methods provide many useful tools for causal inference (88). The most widely adopted one is Mendelian randomization (MR) (89), where genetic variants are used as instrumental variables (IVs). To be a validate IV, a genetic variant (usually a SNP) must satisfy three assumptions (Fig 6.1): Relevance, exchangeability and exclusion restriction (88). In some literature, the exchangeability and exclusion restriction assumptions are combined and referred as 'no direct effect' or 'no horizontal pleiotropy' assumption, where direct effect and horizontal pleiotropy means the variant affects the outcome through another pathway other than mediated by the exposure. Under the above three assumptions, the effect of the variant on the outcome β_{GY} results only from the indirect effect mediated by exposure, which equals to $\beta_{GX} \times \beta_{XY}$, where β_{GX} is the effect of the variant on exposure, and β_{XY} is the effect of the exposure on the outcome. Then we can infer β_{XY} by estimating β_{GY}/β_{GX}.

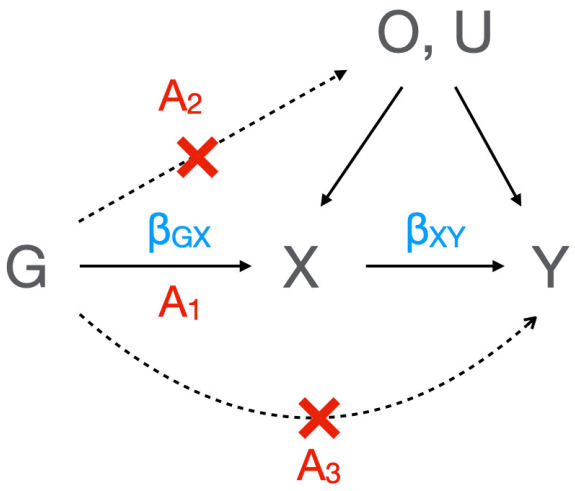

Figure 6.1: Directed acyclic graph and three assumptions in MR. Relevance (A1), the variant (G) associates with the exposure (X); Exchangeability (A2), the variant is independent to all observed (O) and unobserved confounders (U); Exclusion restriction (A3), conditional on exposure and confounders, the variant is independent to the outcome (Y).

Although MR is straightforward to be understood and used, all of the three assumptions can be questionable. The relevance assumption is likely to be violated because the association between a single SNP and the exposure is usually weak; the exchangeability and exclusion restriction can be easily violated due to widespread pleiotropy (90). In practice, the assumptions are mostly untestable, which undermines the reliability of MR results.

To relax these assumptions and to make MR more robust, various methods have been developed based on MR. Most methods only require GWAS summary statistics as input source data. MR with multiple instruments is an essential extension. By analyzing multiple instruments simultaneously, statistical power is increased (91; 92), and more im-

portantly, the 'no horizontal pleiotropy' assumption can be relaxed and tested to some extent. For example, inverse-variance weighted method (93) accepts invalid IVs as long as their direct effects (i) are independent to their effects on the exposure and (ii) have a zero mean; MR-Egger regression (94) further allows the direct effects to have non-zero mean; Median-based methods allow some invalid IVs as long as at least 50% of the IVs are valid (95); Mode-based methods give a consistent estimate when the invalid IVs have heterogeneous effect estimates (96); MR-PRESSO (90) can be used to detect horizontal pleiotropy. The other efforts include two-sample MR (97) where the GWAS summary statistics for exposure and outcome are from separate samples; Multivariable MR (98) includes several variables as exposures to directly model possible pleiotropic pathways, and bidirectional MR (99) aims to infer the direction of causation.

For each new MR-related method, although it relaxes MR assumptions to some extent, it often introduces new untestable assumptions. Therefore in MR and causal inference literature, the importance of triangulation has been frequently emphasized (100; 101; 88). Triangulation refers to the use of multiple approaches, preferably with distinct assumptions, to address the same question. If multiple methods agree on the same conclusion, then the evidence is very strong, especially when the methods are distinct. In **Study IV**, we introduced a method, fundamentally distinct from MR approaches, called Genetic Correlation Contrast for Causality (G3C) to test the existence and direction of a causal relationship between two phenotypes. G3C makes causal inference based on a new theory on genetic correlation heterogeneity. Therefore it uses GWAS results across all the SNPs and does not rely on the 'no horizontal pleiotropy' assumption. These make G3C a useful method for triangulation in causal inference using observational data, particularly GWAS data. G3C can be implemented based solely on GWAS summary statistics, using HDL, the method we developed in **Study II**, to estimate genetic correlations accurately.

Chapter 7

Summary of studies

Data sources

Individual-level data:

The Swedish Twin Registry (TwinGene): Approximately 10,000 genotyped individuals were deep phenotyped. There are 644,556 directly genotyped SNPs, which are imputed to 2,585,290 SNPs using Hapmap 2 build 36.

1000 Genomes: The 1000 Genomes Project applied whole-genome sequencing to a diverse set of individuals from multiple populations. The 1000 Genomes phases 3 data reconstructed the genomes of 2,504 individuals from 26 populations, including 503 European ancestry samples, using a combination of low-coverage whole-genome sequencing, deep exome sequencing, and dense microarray genotyping. The data are publicly available.

UK Biobank: Participants were recruited from the general UK population across 22 centers between 2006-2010 (102). Subjects were aged 40-69 at baseline, underwent extensive phenotyping by questionnaire and clinic measurements, and provided a blood sample. In total 502,664 subjects had been genotyped on an Affymetrix chip including 800,000 variants, that had already been imputed to millions of variants. These individuals had complete phenotyping where various measurements and disease outcomes are included.

Summary-level data:

UK Biobank GWAS round 2 by Neale's group: GWAS on ∼336,000 unrelated individuals of British ancestry for over 2,000 of the available phenotypes in UK Biobank. Age, age^2, inferred sex, age $\times$ inferred sex, age^2 $\times$ inferred sex, and PCs 1-20 were adjusted.

GIANT consortium: The Genetic Investigation of ANthropometric Traits (GIANT) consortium performed GWAS meta-analysis on several anthropometric traits. We used results for six anthropometric traits: BMI, height, weight, hip circumference (HIP), waist circumference (WC) and waist-to-hip ratio (WHR). BMI data are from Locke et al. (103); height data are from Wood et al. (104); weight data are from Randall et al. (105); HIP, WC and WHR data are from Shungin et al. (9).

Study I

Title:

A selection operator for summary association statistics reveals allelic heterogeneity of complex traits

Background:

Many loci mapped in GWAS harbour multiple causal variants. Conditional and joint multi-variant analysis (GCTA-COJO) has been widely used in discovering additional association signals within detected loci. There is theoretical and empirical evidence that the least absolute shrinkage and selection operator (LASSO) provides a better variable selection procedure than stepwise selection procedures, which is implemented in GCTA-COJO.

Aims:

We aimed to obtain LASSO result using GWAS summary statistics (SOJO), and implement SOJO on genomic loci discovered in standard GWAS, for fine-mapping purpose.

Results:

We mathematically proved that LASSO can be achieved using summary-level data and is approximately equivalent to LASSO based on individual-level data. We showed that SOJO provides better sensitivity and specificity in evaluating the number of variants with independent effects in each locus. SOJO suggests causal variants which may be missed by GCTA-COJO. According to our empirical results, human height is not only a highly polygenic trait but also has high allelic heterogeneity within its established hundreds of loci. We built an R package *sojo*, which is available in R-Forge (`https://r-forge.r-project.org/R/?group_id=2030`). A help document can be found at `https://github.com/zhenin/sojo`.

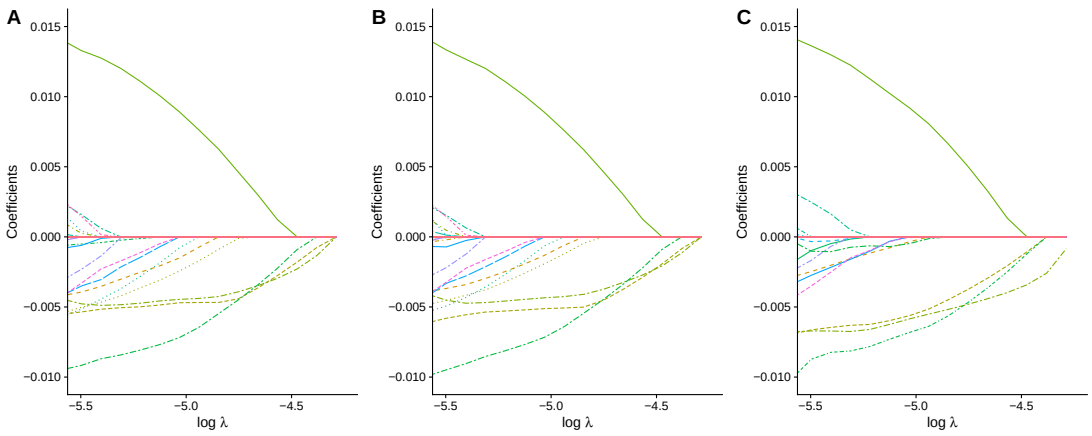

Figure 7.1: An example showing the approximation of SOJO to individual-level LASSO. The GWAS are based on 120,086 individuals in the UKB. The curves represent regularization paths of LASSO coefficients. (A) LASSO based on individual-level data. (B) SOJO based on GWAS summary statistics and LD correlations estimated from the GWAS sample. (C) SOJO based on GWAS summary statistics and LD correlations estimated from 9,617 TwinGene individuals.

Study II

Title:

High-definition likelihood inference of genetic correlations across human complex traits

Background:

Genetic correlation is a key parameter in genetics and genetic epidemiology. Historically, genetic correlation was estimated by using family design data. Currently, when SNP data became available, GREML was used to estimate genetic correlation from LMM, where individual-level genotype data were needed. To estimate genetic correlation from summary-level data, LDSC was developed based on the same LMM as in GREML. However, standard errors of genetic correlation estimates by LDSC were observed to be substantially larger than those of GREML.

Aims:

We aimed to estimate genetic correlations with a full likelihood-based method using summary-level data.

Results:

We noticed that LDSC only partially uses LD information in the modelling of summary association statistics. We developed HDL, a likelihood-based method which is a natural extension of LDSC but fully accounting for LD information. We compared HDL and LDSC in both simulation and application on UKB. Comparing to LDSC, HDL reduces the variance of a genetic correlation estimate by about 60%, which is equivalent to a 2.5-fold increase in sample size. Therefore HDL has higher statistical power to detect genetic correlations between phenotypes. We developed an R package *HDL*, which is hosted on GitHub (`https://github.com/zhenin/HDL`).

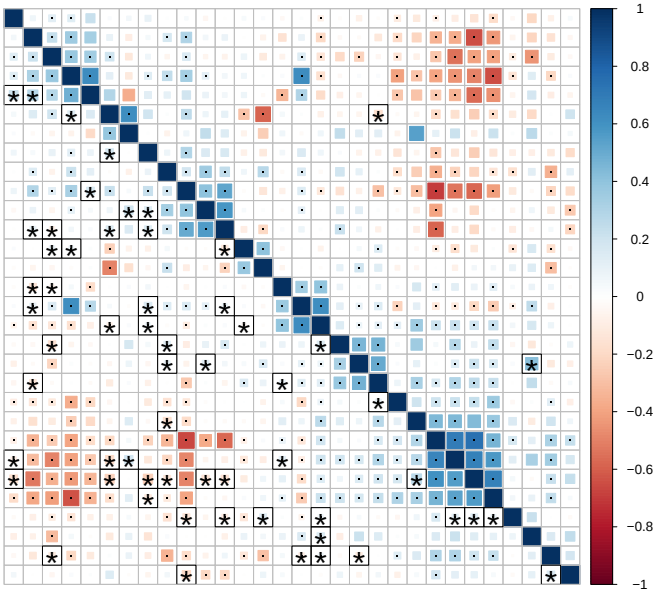

Figure 7.2: The power gain of HDL on genetic correlation estimates over LDSC. The complete version of this plot with trait names can be found in Fig. 2 of **Study II**. Genetic correlation estimates that were significantly different from zero in only one method are marked with asterisk and black square. Lower triangle: HDL estimates; upper triangle: LDSC estimates.

Study III

Title:
Nontrivial replication of loci detected by multi-trait methods

Background:
Various multi-trait methods have been used to obtain higher discovery power than univariate GWAS. However, the replication strategy for SNPs discovered by multi-trait methods has yet to be agreed upon.

Aims:
We aimed to develop more rigorous replication strategies for loci detected by multi-trait methods.

Results:
We introduced Monte-Carlo based correlation methods for evaluating the consistency of multi-trait genetic associations between the discovery and replication samples. By simulations under different scenarios, we illustrated the replication strength of four methods: (i) MANOVA, as a representative of multivariate methods providing unsigned omnibus p-values; (ii) phenotype score replication; (iii) Pearson correlation method; and (iv) Kendall correlation method. Implementation of these methods only requires summary association statistics. Then we proposed a four-level replication strategy for replication of loci detected by multi-trait methods. To demonstrate the application of the four-level replication strategy, we studied the SNPs discovered by MANOVA using GWAS summary statistics from the GIANT consortium as an example and replicated them in UKB.

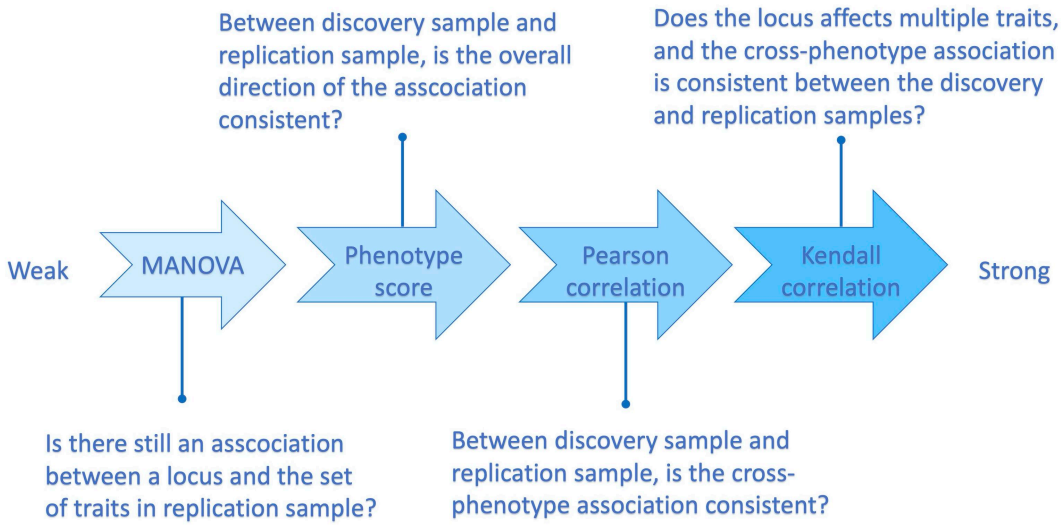

Figure 7.3: Flowchart illustrating the four methods and the questions they can answer.

Study IV

Title:
Inferring causation from heterogeneity in genetic correlations of complex traits

Background:
Causal inference plays an essential role in epidemiology and biomedical science. The gold standard method RCT is powerful but is often infeasible due to cost or ethical and practical concerns. The emerging genetic data popularise MR, where genetic variants are used as IVs. Although MR is easy to use, it relies on the 'no horizontal pleiotropy' assumption. This assumption is often violated in real analysis, especially when the IVs are selected through blind screening from GWAS, and there is a genetic correlation between exposure and outcome.

Aims:
We aimed to develop and implement a test, which is free from the no horizontal pleiotropy assumption, for the existence and direction of a causal relationship between two phenotypes.

Results:
We developed G3C to test for causal inference using GWAS summary statistics. G3C does not rely on the no horizontal pleiotropy assumption, which makes it take full advantage of GWAS results and account for underlying genetic correlation between complex traits. We applied G3C on UKB and discovered new causalities. We performed two-sample MR analysis on G3C discoveries and found significant consistency between the two methods.

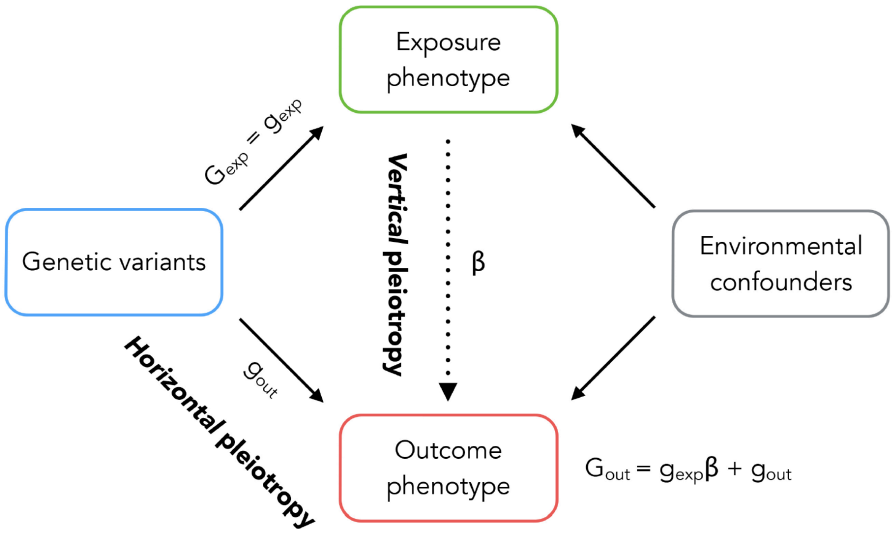

Figure 7.4: Directed acyclic graph shows the parameterization for G3C. g_{exp} and g_{out} are modelled as random effects. The inference is based on the heterogeneity in genetic correlation $r_G = \text{cor}(G_{exp}, G_{out})$ between different samples.

Chapter 8

Future directions

Because of the advances in biotechnologies, biological data collection has been becoming faster, cheaper and more accessible. Not only for genotypes and phenotypes but also for gene expression and different omics information in various tissues and cell types, the amount of multi-level biological data are increasing rapidly. Therefore, in the future of GWAS, we can foresee that GWAS will be done (i) for more phenotypes, (ii) with more individuals, (iii) for different levels of omics data in different tissues, and (iv) in different ethnic groups.

The growing number of individuals in GWAS highlights the advantage of summary-level methods: they are computationally fast and can use available data without pooling individual information. The major challenge, also the opportunity of summary-level methods would be the integration of GWAS results from different levels of biological data. As an example, stratified LDSC (106) incorporates functional annotation into LDSC and enables heritability enrichment analysis. A similar feature can be naturally added to HDL.

Some interesting attempts have been made by using heritability enrichment analysis. For example, Bryois et al. identified cell types underlying brain complex traits by integrating GWAS results with single-cell transcriptomic data (107). However, it is usually difficult to evaluate the correctness of results from such complex analysis. Therefore methods based on distinct assumptions, models, biological levels of data and ethnics could be essential for triangulation purposes. As an example of triangulation, colocalization has been successfully applied in many fine-mapping studies (25).

LD information is crucial for summary-level methods. In most summary-level methods, a reference sample with individual-level genotype data is needed for estimating LD. The estimated LD will then be used to approximate the LD in the GWAS sample. Because a reference sample with larger sample size generates more accurate LD estimates, large publicly available cohorts such as UK Biobank are still in great need. If accurate LD information becomes available for a wider range of ancestries, summary-level methods can be more widely used.

In summary, we are in an era where it is easier to generate data than to interpret them. More methodology and analysis efforts are needed so that we can turn piles of data into useful biological knowledge.

References

[1] Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nature Genetics. 2003;33:228.

[2] Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. The American Journal of Human Genetics. 2012;90(1):7–24.

[3] Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 1996;273(5281):1516–1517.

[4] Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005;308(5720):385–389.

[5] Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. 10 years of GWAS discovery: biology, function, and translation. The American Journal of Human Genetics. 2017;101(1):5–22.

[6] Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Research. 2019;47(D1):D1005–D1012.

[7] Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nature Genetics. 2014;46(11):1173–1186.

[8] Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;518(7538):197–206.

[9] Shungin D, Winkler TW, Croteau-Chonka DC, Ferreira T, Locke AE, Mägi R, et al. New genetic loci link adipose and insulin biology to body fat distribution. Nature. 2015;518(7538):187.

[10] Visscher PM. Sizing up human height variation. Nature Genetics. 2008;40(5):489.

[11] Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, Manolio T, et al. Abundant pleiotropy in human complex diseases and traits. The American Journal of Human Genetics. 2011;89(5):607–618.

[12] Shen X, Klarić L, Sharapov S, Mangino M, Ning Z, Wu D, et al. Multivariate discovery and replication of five novel loci associated with Immunoglobulin G *N*-glycosylation. Nature Communications. 2017;8.

[13] Parkes M, Cortes A, Van Heel DA, Brown MA. Genetic insights into common pathways and complex relationships among immune-mediated diseases. Nature Reviews Genetics. 2013;14(9):661.

[14] McClellan J, King MC. Genetic heterogeneity in human disease. Cell. 2010;141(2):210–217.

[15] Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, Gómez-Marín C, et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature. 2014;507(7492):371.

[16] Evangelou E, Ioannidis JP. Meta-analysis methods for genome-wide association studies and beyond. Nature Reviews Genetics. 2013;14(6):379.

[17] Zeggini E, Ioannidis JP. Meta-analysis in genome-wide association studies. Future Medicine. 2009;.

[18] Ioannidis JP, Patsopoulos NA, Evangelou E. Heterogeneity in meta-analyses of genome-wide association investigations. PLOS ONE. 2007;2(9):e841.

[19] DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials. 1986;7(3):177–188.

[20] Pereira TV, Patsopoulos NA, Salanti G, Ioannidis JP. Discovery properties of genome-wide association signals from cumulatively combined data sets. American Journal of Epidemiology. 2009;170(10):1197–1206.

[21] Ding K, Kullo IJ. Methods for the selection of tagging SNPs: a comparison of tagging efficiency and performance. European Journal of Human Genetics. 2007;15(2):228–236.

[22] MacArthur D, Manolio T, Dimmock D, Rehm H, Shendure J, Abecasis G, et al. Guidelines for investigating causality of sequence variants in human disease. Nature. 2014;508(7497):469–476.

[23] Yang J, Ferreira T, Morris AP, Medland SE, Madden PA, Heath AC, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nature Genetics. 2012;44(4):369–375.

[24] Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E. Identifying causal variants at loci with multiple signals of association. Genetics. 2014;198(2):497–508.

[25] Hormozdiari F, Van De Bunt M, Segre AV, Li X, Joo JWJ, Bilow M, et al. Colocalization of GWAS and eQTL signals detects target genes. The American Journal of Human Genetics. 2016;99(6):1245–1260.

[26] Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. Nature Genetics. 2016;48(5):481.

[27] Ongen H, Dermitzakis ET. Alternative splicing QTLs in European and African populations. The American Journal of Human Genetics. 2015;97(4):567–575.

[28] Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics. 2018;19(8):491–504.

[29] Estivill X, Bancells C, Ramos C. Geographic distribution and regional origin of 272 cystic fibrosis mutations in European populations. Human Mutation. 1997;10(2):135–154.

[30] Hormozdiari F, Zhu A, Kichaev G, Ju CJT, Segrè AV, Joo JWJ, et al. Widespread allelic heterogeneity in complex traits. The American Journal of Human Genetics. 2017;100(5):789–802.

[31] Gaulton KJ, Ferreira T, Lee Y, Raimondo A, Mägi R, Reschen ME, et al. Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci. Nature Genetics. 2015;47(12):1415–1425.

[32] Furnival GM, Wilson RW. Regressions by leaps and bounds. Technometrics. 1974;16(4):499–511.

[33] Efroymson M. Multiple regression analysis. Mathematical Methods for Digital Computers. 1960;p. 191–203.

[34] Efron B, Hastie T, Johnstone I, Tibshirani R, et al. Least angle regression. The Annals of Statistics. 2004;32(2):407–499.

[35] Derksen S, Keselman HJ. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. British Journal of Mathematical and Statistical Psychology. 1992;45(2):265–282.

[36] Bøvelstad HM, Nygård S, Størvold HL, Aldrin M, Borgan Ø, Frigessi A, et al. Predicting survival from microarray data—a comparative study. Bioinformatics. 2007;23(16):2080–2087.

[37] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996;p. 267–288.

[38] Omranian N, Eloundou-Mbebi JM, Mueller-Roeber B, Nikoloski Z. Gene regulatory network inference using fused LASSO on multiple data sets. Scientific Reports. 2016;6.

[39] Yang AY, Sastry SS, Ganesh A, Ma Y. Fast 1-minimization algorithms and an application in robust face recognition: A review. In: 2010 IEEE International Conference on Image Processing. IEEE; 2010. p. 1849–1852.

[40] Tibshirani RJ, Taylor J, et al. The solution path of the generalized lasso. The Annals of Statistics. 2011;39(3):1335–1371.

[41] Heinig M, Petretto E, Wallace C, Bottolo L, Rotival M, Lu H, et al. A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. Nature. 2010;467(7314):460–464.

[42] Barrett JC, Clayton DG, Concannon P, Akolkar B, Cooper JD, Erlich HA, et al. Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nature Genetics. 2009;41(6):703–707.

[43] Ning Z, Lee Y, Joshi PK, Wilson JF, Pawitan Y, Shen X. A selection operator for summary association statistics reveals allelic heterogeneity of complex traits. The American Journal of Human Genetics. 2017;101(6):903–912.

[44] Guan Y, Stephens M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. The Annals of Applied Statistics. 2011;p. 1780–1815.

[45] Chen W, Larrabee BR, Ovsyannikova IG, Kennedy RB, Haralambieva IH, Poland GA, et al. Fine mapping causal variants with an approximate Bayesian method using marginal test statistics. Genetics. 2015;200(3):719–736.

[46] Stearns FW. One hundred years of pleiotropy: a retrospective. Genetics. 2010;186(3):767–773.

[47] Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al. An atlas of genetic correlations across human diseases and traits. Nature Genetics. 2015;47(11):1236.

[48] Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Human Molecular Genetics. 2014;23(R1):R89–R98.

[49] Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. Nature Genetics. 2018;p. 1.

[50] Grotzinger AD, Rhemtulla M, de Vlaming R, Ritchie SJ, Mallard TT, Hill WD, et al. Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. Nature Human Behaviour. 2019;3(5):513–525.

[51] Tenesa A, Haley CS. The heritability of human disease: estimation, uses and abuses. Nature Reviews Genetics. 2013;14(2):139–149.

[52] Lichtenstein P, Yip BH, Björk C, Pawitan Y, Cannon TD, Sullivan PF, et al. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. The Lancet. 2009;373(9659):234–239.

[53] Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. The American Journal of Human Genetics. 2011;88(1):76–82.

[54] Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. Bioinformatics. 2012;28(19):2540–2542.

[55] Lee SH, Ripke S, Neale BM, Faraone SV, Purcell SM, Perlis RH, et al. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. Nature Genetics. 2013;45(9):984.

[56] Liu JZ, Van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nature Genetics. 2015;47(9):979–986.

[57] Fritsche LG, Igl W, Bailey JNC, Grassmann F, Sengupta S, Bragg-Gresham JL, et al. A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. Nature Genetics. 2016;48(2):134–143.

[58] Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nature Methods. 2014;11(4):407–409.

[59] Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Salem RM, et al. Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nature Genetics. 2015;47(3):284.

[60] Lee SH, Van der Werf JH. MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. Bioinformatics. 2016;32(9):1420–1422.

[61] Yang J, Weedon MN, Purcell S, Lettre G, Estrada K, Willer CJ, et al. Genomic inflation factors under polygenic inheritance. European Journal of Human Genetics. 2011;19(7):807.

[62] Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature Genetics. 2015;47(3):291–295.

[63] Okbay A, Beauchamp JP, Fontana MA, Lee JJ, Pers TH, Rietveld CA, et al. Genome-wide association study identifies 74 loci associated with educational attainment. Nature. 2016;533(7604):539.

[64] Robinson EB, St Pourcain B, Anttila V, Kosmicki JA, Bulik-Sullivan B, Grove J, et al. Genetic risk for autism spectrum disorders and neuropsychiatric variation in the general population. Nature Genetics. 2016;48(5):552.

[65] Zheng J, Erzurumluoglu AM, Elsworth BL, Kemp JP, Howe L, Haycock PC, et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. Bioinformatics. 2017;33(2):272–279.

[66] Van Rheenen W, Peyrot WJ, Schork AJ, Lee SH, Wray NR. Genetic correlations of polygenic disease traits: from theory to practice. Nature Reviews Genetics. 2019;20(10):567–581.

[67] Speed D, Balding DJ. SumHer better estimates the SNP heritability of complex traits from summary statistics. Nature Genetics. 2019;51(2):277–284.

[68] Weissbrod O, Flint J, Rosset S. Estimating SNP-based heritability and genetic correlation in case-control studies directly and with summary statistics. The American Journal of Human Genetics. 2018;103(1):89–99.

[69] Ning Z, Pawitan Y, Shen X. High-definition likelihood inference of genetic correlations across human complex traits. Nature Genetics. 2020;52:859-–-864.

[70] Van der Sluis S, Posthuma D, Dolan CV. TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. PLOS Genetics. 2013;9(1):e1003235.

[71] Ferreira MA, Purcell SM. A multivariate test of association. Bioinformatics. 2008;25(1):132–133.

[72] Aschard H, Vilhjálmsson BJ, Greliche N, Morange PE, Trégouët DA, Kraft P. Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. The American Journal of Human Genetics. 2014;94(5):662–676.

[73] Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. Pleiotropy in complex traits: challenges and strategies. Nature Reviews Genetics. 2013;14(7):483.

[74] Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, Wallace C, et al. Pervasive sharing of genetic effects in autoimmune disease. PLOS Genetics. 2011;7(8):e1002254.

[75] van der Sluis S, Posthuma D, Dolan CV. TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. PLOS Genetics. 2013;9(1):e1003235.

[76] Kim J, Bai Y, Pan W. An adaptive association test for multiple phenotypes with GWAS summary statistics. Genetic Epidemiology. 2015;39(8):651–663.

[77] Stephens M. A unified framework for association analysis with multiple related phenotypes. PLOS ONE. 2013;8(7):e65245.

[78] Zhu X, Feng T, Tayo BO, Liang J, Young JH, Franceschini N, et al. Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. The American Journal of Human Genetics. 2015 Jan;96(1):21–36.

[79] Park H, Li X, Song YE, He KY, Zhu X. Multivariate analysis of anthropometric traits using summary statistics of genome-wide association studies from GIANT consortium. PLOS ONE. 2016;11(10):e0163912.

[80] Porter HF, O Reilly PF. Multivariate simulation framework reveals performance of multi-trait GWAS methods. Scientific Reports. 2017;7:38837.

[81] Liang J, Le TH, Edwards DRV, Tayo BO, Gaulton KJ, Smith JA, et al. Single-trait and multi-trait genome-wide association analyses identify novel loci for blood pressure in African-ancestry populations. PLOS Genetics. 2017;13(5):e1006728.

[82] Gialluisi A, Andlauer TF, Mirza-Schreiber N, Moll K, Becker J, Hoffmann P, et al. Genome-wide association scan identifies new variants associated with a cognitive predictor of dyslexia. Translational Psychiatry. 2019;9(1):1–15.

[83] Karnes JH, Bastarache L, Shaffer CM, Gaudieri S, Xu Y, Glazer AM, et al. Phenome-wide scanning identifies multiple diseases and disease severity phenotypes associated with HLA variants. Science Translational Medicine. 2017;9(389):eaai8708.

[84] Luo L, Shen J, Zhang H, Chhibber A, Mehrotra DV, Tang ZZ. Multi-trait analysis of rare-variant association summary statistics using MTAR. Nature Communications. 2020;11(1):1–11.

[85] Zhernakova A, Van Diemen CC, Wijmenga C. Detecting shared pathogenesis from the shared genetics of immune-related diseases. Nature Reviews Genetics. 2009;10(1):43.

[86] Gaziano JM. Vitamin E and cardiovascular disease: observational studies. Annals of the New York Academy of Sciences. 2004;1031(1):280–291.

[87] Hernán MA. A definition of causal effect for epidemiological research. Journal of Epidemiology & Community Health. 2004;58(4):265–271.

[88] Pingault JB, O' reilly PF, Schoeler T, Ploubidis GB, Rijsdijk F, Dudbridge F. Using genetic data to strengthen causal inference in observational research. Nature Reviews Genetics. 2018;19(9):566–580.

[89] Davey Smith G, Ebrahim S. 'Mendelian randomization': can Genetic Epidemiology contribute to understanding environmental determinants of disease? International Journal of Epidemiology. 2003;32(1):1–22.

[90] Verbanck M, Chen Cy, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. Nature Genetics. 2018;50(5):693–698.

[91] Palmer TM, Lawlor DA, Harbord RM, Sheehan NA, Tobias JH, Timpson NJ, et al. Using multiple genetic variants as instrumental variables for modifiable risk factors. Statistical Methods in Medical Research. 2012;21(3):223–242.

[92] Brion MJA, Benyamin B, Visscher PM, Smith GD. Beyond the single SNP: emerging developments in Mendelian randomization in the "Omics"era. Current Epidemiology Reports. 2014;1(4):228–236.

[93] Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. Genetic Epidemiology. 2013;37(7):658–665.

[94] Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. International Journal of Epidemiology. 2015;44(2):512–525.

[95] Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent estimation in Mendelian randomization with some invalid instruments using a weighted median estimator. Genetic Epidemiology. 2016;40(4):304–314.

[96] Hartwig FP, Davey Smith G, Bowden J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. International Journal of Epidemiology. 2017;46(6):1985–1998.

[97] Bowden J, Del Greco M F, Minelli C, Davey Smith G, Sheehan N, Thompson J. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. Statistics in Medicine. 2017;36(11):1783–1802.

[98] Burgess S, Thompson SG. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. American Journal of Epidemiology. 2015;181(4):251–260.

[99] Gage SH, Jones HJ, Burgess S, Bowden J, Smith GD, Zammit S, et al. Assessing causality in associations between cannabis use and schizophrenia risk: a two-sample Mendelian randomization study. Psychological Medicine. 2017;47(5):971–980.

[100] Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. International Journal of Epidemiology. 2016;45(6):1866–1886.

[101] Munafò MR, Smith GD. Robust research needs many lines of evidence. Nature. 2018;.

[102] Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: an Open Access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLOS Medicine. 2015;12(3):1–10.

[103] Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;518(7538):197–206.

[104] Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nature Genetics. 2014 Nov;46(11):1173–1186.

[105] Randall JC, Winkler TW, Kutalik Z, Berndt SI, Jackson AU, Monda KL, et al. Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. PLOS Genetics. 2013;9(6):e1003500.

[106] Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nature Genetics. 2015;47(11):1228.

[107] Bryois J, Skene NG, Hansen TF, Kogelman LJ, Watson HJ, Liu Z, et al. Genetic identification of cell types underlying brain complex traits yields insights into the etiology of Parkinson' s disease. Nature Genetics. 2020;52(5):482–493.

Acknowledgements

It has been such a wonderful journey. The moment my hands touched the keyboard to draft this section, all of those lovely scenes came back to me like a movie. I would like to thank all of you who made my past years as a PhD student into such a joyful experience.

Thank you **Xia Shen**, for taking me on as your first PhD student. It is so inspiring and fun to do research with you. Throughout my PhD studies, you have been a very positive, passionate, creative, and supportive supervisor. There were countless moments during my PhD when I felt my projects were hopeless and frustrating. But you were always there to encourage and guide me to find a path together with you. Although for the most time you were not physically around me, you were always available for taking any question from me, which is much more important. "Hypothesize boldly, while prove it carefully", I will keep it in mind forever. Additionally, you are a great friend. From sleeping on your sofa on the first day of my PhD student life, to moving to Upplands Väsby and being your neighbor, I feel so fortunate to have you as my supervisor and friend.

Thank you **Yudi Pawitan**, for being my co-supervisor, and for teaching me how to be a sensible person. Your deep insight into statistics and genetics has enlightened me so much. But what I have learned from you is far more than knowledge. All of your sharp questions, critical comments, tricky puzzles, friendly advice, and patient explanations drive me forward from different perspectives. Besides research, your passion in life deeply impressed and touched me too. Food, music, bet, books, storm names... all of these fun discussions, together with those dinners you have treated me, made my PhD super enjoyable and memorable.

Thank you **Gunilla Sonnebring**, my dear mentor, for all of the lovely conversations and lunches. When I am delighted, I can share the happiness with you; when I am down, I can always bother you to get encouragement and a big hug. Additionally, I want to thank you for recruiting me into the Praesepe editorial board. By the eight passages I have "published" on Praesepe, I shared fun stories that happened in my life and improved my writing skill. Also, I would like to thank you for being my "cooking supervisor". All of the projects you recommended to me were highly satisfying. They made me better mentally and physically.

Thank you **Gabriel Isheden**, I feel super fortunate to be your office-mate. You are such an optimistic person, which makes our office a "happy room". We have had so many wonderful discussions on everything, and you are always generous to teach me Swedish and investment (not Brazilian jiu-jitsu fortunately). Thank you **Lu Pan**, for joining our office and bringing sweet ideas such as daily compliments. Of course, I will never forget my previous office 5134. Although it located at the corner of the basement floor, thanks to **Jiayao Lei** and **Qing Shen**, it became an awesome office. Thank you **Chen Wang**, our friendship started at office 5134 and will continue for life. Looking forward to keeping chatting with you about games, dieting, medical science, statistics, and hiking (the order is completely random). Wish you and **Xinge Li** keep living a happy life!

To **Nghia Trung Vu**, **Stefano Calza**, **Tian Mou**, **Wenjiang Deng**, **Dat T Nguyen**, **Quang Thinh Trac** and **Sarath Kumar Murugan**, thank you for the fruitful discussions during our weekly group meetings, and for the splendid celebration dinners. It has been a great pleasure to work with you. Special thanks to **Minh Khoa Vu**, a talented and generous artist, for letting me use your masterpiece as the cover page of my thesis for free.

Thank you **Youngjo Lee**, for hosting us with great hospitality. The cozy conference room next to your office is definitely a magical place. I want to thank the whole **Wilson group** for the suggestions and comments, and the wonderful retreats. Especially, I want to thank **Linda Repetto** for illuminating questions; and **Paul Timmers** for comments on my HDL package.

Thank you **Marie Jansson**, **Alessandra Nanni**, **Jacqueline Knight**, **Gunilla Nilsson Roos**, and **Camilla Ahlqvist**, for making my PhD so smooth with your kindness and professional skills.

Thank you **Arvid Sjölander** and **Alex Ploner** for recruiting me as your teaching assistant. I have learned a lot from teaching students. Thank you **Marie Reilly**, not only for recruiting me in your workshop for medical students but for all of the energizing conversations and cupcakes!

One of the best memory I had at MEB is the Harry Potter Christmas party. Thank you **Kathleen Bokenberger**, **Hannah Bower**, **Shuyang Yao**, **Emilio Ugalde Morales** and **Laura Ghirardi** for making the magical dream come true. And **Elisabeth Dahlqwist**, thank you for inviting me to join the Christmas party, and for all of those happy moments during these years.

I would like to thank my brilliant colleagues currently or previously working at the

MEB biostatistics group: **Anna Johansson**, **Robert Karlsson**, **Li Yin**, **Cecilia Lundholm**, **Agnieszka Szwajda**, **Mikael Andersson Franko**, **Henric Winell**, **Keith Humphreys**, **Maya Alsheh Ali**, **Rickard Strandberg**, **Benjamin Christoffersen**, **Linda Abrahamsson**, **Iuliana Ciocanea-Teodorescu**, **Erin Gabriel**, **Pablo Gonzalez Ginestet**, **Alessandro Gasparini**, **Mark Clements**, **Shuang Hao**, **Andreas Karlsson**, **Xingrong Liu**, **Lili Meng**, **Yingying Yang**, **Anastasia Lam**, **Aminata Ndiaye**, **Bénédicte Delcoigne**, **Sophie Debonneville**, **Zhipeng Wang**, **Tingyou Zhou**, **Dan Bai**, **Paul Dickman**, **Therese Andersson**, **Yuliya Leontyeva**, **Caroline Weibull**, **Nurgul Batyrbekova**, **Nikolaos Skourlis**, **Paul Lambert**, **Martin Eklund**, **Thorgerdur Palsdottir**, **Henrik Olsson**, **Peter Ström**, **Rino Bellocco**, **Alessandra Grotta**, **Daniela Mariosa**, **Alessio Crippa**, **Marco Trevisan**, **Juni Palmgren**, **Sven Sandin** and **Johan Zetterqvist**. Your high quality works make the biostatistics group keep being energetic and productive.

Thank you **Shuyang Yao**, **Yunzhang Wang** and **Ruyue Zhang**, for kindly accepting to be committees in my pre-dissertation. And **Patrik Magnusson**, thank you for being the chairperson of my defense, for your fantastic show at Christmas party, and for keeping trying to teach me Swedish.

To **Jiangwei Sun**, **Lin Li**, **Yunzhang Wang**, **Yinxi Wang**, **Xia Li** and **Weiwei Bian**, thank you for your fabulous work on our book! Hopefully, it can be published soon and can be helpful for a wide range of readers.

The "MEB spirit" comes from brilliant MEBers. Thanks to the other past and present MEBers including but not excluded to: **Tong Gong**, **Jie Song**, **Vivekananda Lanka**, **Yiqiang Zhan**, **Ruoqing Chen**, **Yi Lu**, **Wei He**, **Zheng Chang**, **Donghao Lu**, **Xu Chen**, **Jiangrong Wang**, **Bojing Liu**, **Haomin Yang**, **Tingting Huang**, **Xiaoying Kang**, **Qi Chen**, **Julien Bryois**, **Shihua Sun**, **Puisan Tan**, **Jonas Ludvigsson**, **Le Zhang**, **Felix Grassmann**, **Ruyue Zhang**, **Erika Nordenhagen**, **Jingru Yu**, **Can Cui**, **Xinhe Mao**, **Yang Xu**, **Erwei Zeng**, **Shengxin Liu**, **Qian Yang**, **Guobin Su**, **Cen Chen**, **Fei He** and **Kristina Johnell**.

To my lunch gang members: **Di Wu**, **Xingwu Zhou**, **Xiaoyan Qian** and **Xiang Jiao**. I learned so much about science and life from our discussions or arguments. Thank you for sharing your stories and ideas.

I would never have become a PhD in Sweden without the guidance from **Fan Yang Wallentin**. Thanks to the exchange program you organized, I got to know Sweden; thanks to your suggestion on the second year of my master, I got to know Yudi and Xia. I also want to thank **Shaobo Jin & Xuan Li**, and **Zhizheng Wang & Yi Ding** for your company and the dragon ball nights. And **Fangkai Yang**, my "oldest" friend in Sweden,

wish you all the best wherever you go after your PhD.

亲爱的爸妈，感谢你们对我的爱，鼓励与支持。转瞬之间，离家已十年。我一定会继续努力，不辜负你们对我的教导与期待。感谢亲人们对我的关心与照顾。对于在外漂泊的我，你们是我永久的港湾。

Finally, my **Ximeng**, I became a better person because of you. Thank you for everything. I love you forever.

I always consider myself as a lucky one. Not because I keep winning lotteries, but because I have met all of you. Thank you!